Hire a Linguist!: Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions

Kexun Zhang[1] **Yee Man Choi**[1] **Zhenqiao Song**[1] **Taiqi He**[1]
William Yang Wang[2] **Lei Li**[1]
[1]Carnegie Mellon University [2]UC Santa Barbara
{kexunz,yeemanc,zhenqias,taiqih}@andrew.cmu.edu
william@ucsb.edu leili@cs.cmu.edu

Abstract

How can large language models (LLMs) process and translate endangered languages? Many languages lack a large corpus to train a decent LLM; therefore existing LLMs rarely perform well in unseen, endangered languages. On the contrary, we observe that 2000 endangered languages, though without a large corpus, have a grammar book or a dictionary. We propose LINGOLLM, a training-free approach to enable an LLM to process unseen languages that hardly occur in its pre-training. Our key insight is to demonstrate linguistic knowledge of an unseen language in an LLM's prompt, including a dictionary, a grammar book, and morphologically analyzed input text. We implement LINGOLLM on top of two models, GPT-4 and Mixtral, and evaluate their performance on 5 tasks across 8 endangered or low-resource languages. Our results show that LINGOLLM elevates translation capability from GPT-4's 0 to 10.5 BLEU for 10 language directions. Our findings demonstrate the tremendous value of linguistic knowledge in the age of LLMs for endangered languages. Our data, code, and model generations can be found at https://github.com/LeiLiLab/LingoLLM.

1 Introduction

Large language models (LLMs) are already powerful in many language understanding and generation tasks (Brown et al., 2020; Ouyang et al., 2022). Their language processing capabilities rely on very large amounts of training data (Kaplan et al., 2020; Hoffmann et al., 2022). For example, a recent LLM Llama-2 uses a pre-training dataset with 2 trillion tokens (Touvron et al., 2023). While languages such as English or Spanish enjoy abundant accessible data, the majority of the world's 7000 languages lack a rich corpus, including most endangered languages recognized by UNESCO (Moseley, 2010). Existing LLMs such as Llama (Touvron et al., 2023) and GPT-4 show poor performance on languages that may not occur in pre-training (Robinson et al., 2023). We believe that speakers of endangered languages deserve equitable access to NLP technologies including LLMs. How can we enable an LLM with language processing capabilities on unseen and endangered languages?

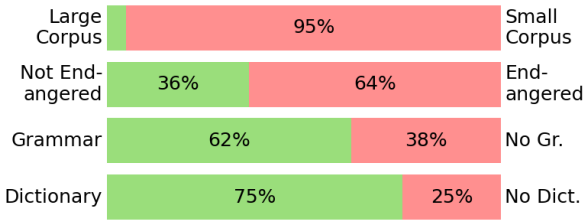

Figure 1: Among the world's ∼7000 languages, 95% don't have enough data (>100K sentences) for training LLMs (Bapna et al., 2022), while most have a grammar book (60%) or dictionary (75%) (Nordhoff and Hammarström, 2011), including many endangered languages (Moseley, 2010). Therefore, we utilize these linguistic descriptions to bring LLMs to endangered languages.

We are motivated by how human linguists analyze utterances in a language they don't know — they use existing grammar books and dictionaries. Fortunately, thanks to the efforts of generations of linguists over the years, many endangered languages have published dictionaries and descriptive grammar. Compared to LLMs' training corpora, which mostly consist of unstructured text, these linguistic descriptions have two major differences. First, they are instructional. Though they are much smaller than typical training sets, they contain explicit grammar rules of a language that can be used as instructions for both LLMs and humans. Second, linguistic descriptions have much broader coverage. As shown in Figure 1, very few languages have training corpora, but most have documented grammar or dictionary. However, directly using these linguistic descriptions in an LLM's prompt is infeasible. A grammar book and a dictionary are often too large to fill in the prompt of an LLM.

Findings of the Association for Computational Linguistics: ACL 2024, pages 15654–15669

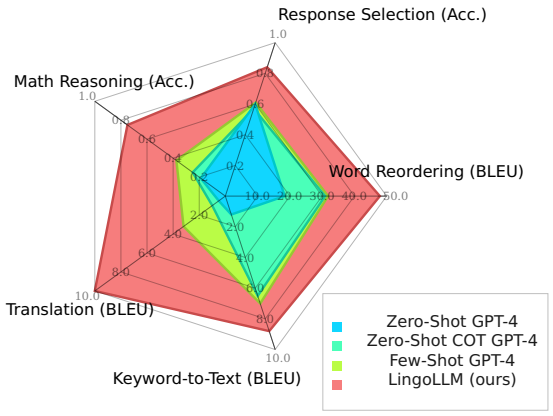

Figure 2: LINGOLLM significantly outperforms GPT-4 on 5 NLP tasks across 8 endangered or low-resource languages.

In this paper, we propose LINGOLLM, an efficient approach to enable an LLM to process and translate unseen languages that never occur in its pre-training. Our key insight is to properly exploit linguistic description of an unseen language, including a dictionary, a grammar book, and morphologically analyzed input text. LINGOLLM first preprocesses input text in an endangered language via a morphological analyzer and a dictionary, both from linguistic descriptions of the language. The inputs, annotated with grammar features and word-level translations, are passed to an LLM along with the grammar book. The LLM then translates the endangered language inputs to a high-resource language like English to process them. LINGOLLM is training-free as it only requires the underlying LLM to be instruction-tuned. LINGOLLM can adapt to languages according to the availability of different types of linguistic descriptions.

We implement LINGOLLM on top of two models, GPT-4 and Mixtral. Our experiments consist of a total of 5 tasks (including translation from/to English, mathematical reasoning, response selection, word reordering, and keyword-to-text) in 8 endangered/low-resource languages that are geographically and typologically diverse. As shown in Figure 2, LINGOLLM significantly improves GPT-4's performance on all 5 tasks by a large margin. Noticeably the translation quality increases from an incomprehensible 0.5 to 10 BLEU points.

Our contributions are:

- We propose LINGOLLM, an approach to integrate linguistic descriptions to process and translate text in endangered languages.
- With the help of linguists, we build processing systems for 8 typologically and geographically diverse endangered or low-resource languages according to the availability of different linguistic descriptions.
- Our experiments show superior performance of LINGOLLM on all tasks, compared to strong baselines (GPT-4 and Mixtral). LINGOLLM elevates translation capability from GPT-4's 0 BLEU to 10.5 BLEU for 10 language directions. It improves GPT-4's mathematical reasoning accuracy from 18% to 75%, and response selection accuracy from 43% to 63%.

2 Related Work

Various recent studies explore the possibility of LLMs for low-resource languages on machine translation (Hendy et al., 2023) and other NLP tasks (Ahuja et al., 2023; Huang et al., 2023). Their scope of evaluation mostly covers languages whose resources are low but still exist. Hence LLMs can still have a non-zero translation ability. We go beyond their scope towards languages that are truly extinct where LLMs' zero-shot translation of them is near *zero*. Moreover, our method relies on external linguistic descriptions rather than internal knowledge of LLMs, focusing on how LLMs can utilize information they don't know instead of how they can "recall" information they have seen in training.

Evaluating LLMs for Low-Resource Languages. Many (Jiao et al., 2023; Hendy et al., 2023; Zhu et al., 2023) suggest that LLMs do poorly on low-resource languages. Robinson et al. (2023) evaluated ChatGPT's machine translation performance on 204 languages and found that ChatGPT consistently underperforms traditional machine translation models on low-resource languages. Ahuja et al. (2023) evaluated several LLMs on 16 NLP tasks and found significant performance drops on low-resource languages. While their conclusions corroborate our motivation, the languages they evaluate are very likely to exist in LLMs' training set, as indicated by LLMs' significantly positive performance in the zero-shot setting. On the contrary, the focus of our paper is on endangered languages that are truly extinct in the training data with near-zero performance from LLMs.

Improving LLMs for Low-Resource Languages. Existing studies improve LLMs' performance on low-resource languages via prompt engineering. Few-shot prompting (Ahuja et al., 2023) puts input-output pairs in the prompt as exemplars using differ-

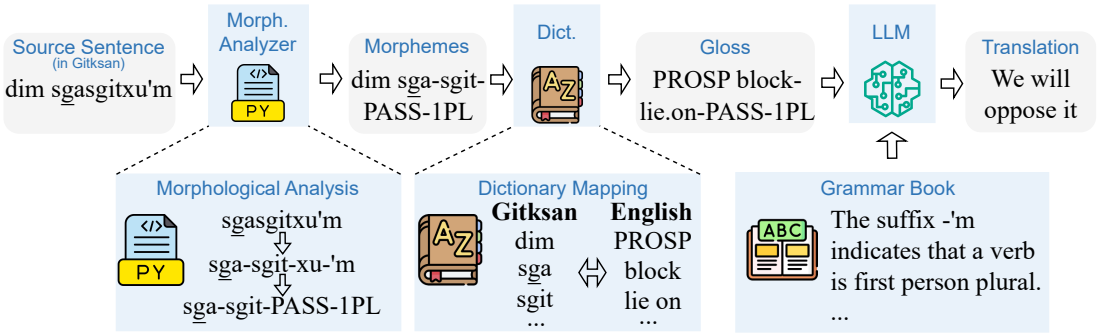

Figure 3: LINGOLLM uses a morphological analyzer to transform the source sentence into morphemes, looks up the morphemes in a dictionary to obtain the gloss, and finally feeds both the gloss and a grammar book to an LLM to obtain the result.

ent selection strategies. Chain-of-thought prompting further prompts the model to solve the problem step by step, either by explicit instruction (Huang et al., 2023) or by in-context examples of step-by-step answers (Shi et al., 2022).

Another line of work relies on external modules and knowledge. Ahuja et al. (2023) uses commercial translation systems to first translate a task to English before giving it to LLMs. Gao et al. (2023) augments the input sequence with POS tagging. However, both commercial translators and POS taggers are hard to find for the languages we evaluate. Our paper is closest to Tanzer et al. (2023), where they also depend on dictionaries and grammar books to translate an endangered language with LLMs. Their main goal was to propose an uncontaminated translation benchmark for evaluation purposes on a single endangered language. Compared to Tanzer et al. (2023), we go beyond machine translation to multiple NLP tasks and evaluate the proposed LINGOLLM extensively on 8 endangered/low-resource languages that are diverse. Also, we make use of morphological analyzers, an extra symbolic module that can be derived from grammar books.

3 The LINGOLLM Approach

Since we do not have enough data to fine-tune an LLM, LINGOLLM equips an LLM with linguistic knowledge to process text in an endangered language (Figure 3). We obtain a morphological analyzer, a dictionary, and a grammar description for each target language. Our method consists of four steps: 1) Given an utterance in an endangered language, we first use a morphological analyzer to split each word into *morphemes*; 2) We search for the closest matches from a dictionary for each morpheme to obtain an annotated *gloss*; 3) We prompt an instruction-tuned LLM with the annotated gloss and a grammar description to get the *translation* in a high-resource language such as English; and 4) We further process the translated text using a same LLM for the downstream task.

To adapt LINGOLLM to a new language, we first collect linguistic descriptions with the help from linguists. We use three types of linguistic descriptions in LINGOLLM – morphological analyzers, dictionaries, and grammar books. For all languages studied in this paper, we look for references to these descriptions on Glottolog[1] and then collect them via web interfaces or ebooks. For one type of description, we build a universal interface to use them despite different underlying formats. We list the collected linguistic descriptions in Appendix B.

3.1 Morphological Analysis: Source Sentence → Morphemes

A morphological analyzer is a program that maps a word to a sequence of morphemes, the smallest meaningful constituents. Morphemes include *stems* that indicate concrete meanings and *linguistic features* that indicate grammatical roles. For example, an English morphological analyzer might map *cats* to *cat +Noun +Plural*, where *cat* is a stem and *+Noun* and *+Plural* are two features. Features make it easier for people and LLMs to find out the grammatical roles of a word, while stems are more convenient for dictionary search than their inflections and derivations.

We use existing finite-state morphological analyzers to identify the stem and features of each word in the source sentence. These analyzers are written as finite-state transducer using the python implementation of foma (Hulden, 2009). We directly apply them to the source words to obtain

[1] https://glottolog.org/

the morphemes. For example, in Figure 3, the Gitksan word *sgasgitxu'm* is transformed into four morphemes *sga-sgi-PASS-1PL*, where the stems are *sga* (meaning "to block") and *sgi* (meaning "to lie on"), and the features indicate that it's a verb with a passive voice whose subject is first person plural.

One concern that may be raised about using morphological analyzers is their availability because they may not exist for as many languages as grammar books do. While there are no statistics about the availability of these analyzers, this issue can be resolved by having trained linguists create morphological analyzers from grammar books.

3.2 Dictionary Mapping: Morphemes → Gloss

We use a dictionary to map the stems (morphemes with concrete meanings) to their dictionary definitions. The word-level translations, along with the stems, constitute the gloss of the source sentence. The gloss can then be utilized by the LLM to formulate sentence-level translation. The dictionary mapping process is not as straightforward as it seems and can involve multiple steps.

Step 0. Normalizing the script. Before we implement the mapping from source words to their translations, we must make sure the scripts used by the dictionary and the input are the same as those in the test set. This is often not the case, especially for endangered languages. For example, the Manchu dictionary we use (Norman, 2020) represents the phonemes /tʃʰ/, /ʃ/, /u/ as q, x, and v, while our Manchu inputs represent the three phonemes as c, š, and ū. We manually compare the written forms of the same words in the input and the dictionary to derive rules that map one script to another.

Step 1. Deciding the input: words or stems? We can either use source words or their stems as the input to the dictionary, depending on the availability of a morphological analyzer and the scope of words in the dictionary. Usually, it is easier to find matches in the dictionary for word stems produced by a morphological analyzer. But we have to use the original words when such an analyzer is not available. Many dictionaries only have entries for one form of a verb. For example, Manchu dictionaries might only contain verbs in their present tense form with the suffix *-mbi*. When this is the case, we have to get the stems of the words first or use some sort of fuzzy matching.

Step 2. Finding the closest match. Online dictionaries' search algorithms often provide multiple possible matches for a word. In the case where we are unable to retrieve the word stem or the word stem does not exist in the dictionary, we would not be able to find an exact match from the dictionary and need to choose the closest match using the edit distance. For instance, the word stem for the Gitksan word *mismaaxwsxum* (a plural marking attribute meaning white) is *mismaaxwsxw*, but we are unable to find *mismaaxwsxw* within the Gitksan dictionary we use (Gitksan Research Lab, 2023). However, we can find the following partial match, *maaxwsxw* or *maxwsxw* meaning "to be white", *maaxws* meaning "snow (on ground)", *misaax* meaning "daylight", and *sawnsxw* meaning "paper". Using the edit distance as a selection metric, we can retrieve the closest matches *maaxwsxw* or *maxwsxw* that are most related to the word *mismaaxwsxum*.

Step 3. Collecting other relevant words. Some dictionaries' entries contain references to other entries. The content of these referenced entries provides complementary information related to the matched word. For example, the entry for *qoohiyan* in our Manchu dictionary states that it stands for "Korea" the place. It also references another Manchu word *solho*, meaning "Korean" the people. To collect such information, we traverse the graph formed by cross-entry links starting from the match until all connected entries are found or the number of found entries exceeds a threshold.

3.3 Incorporating Grammar Knowledge Gloss → Translation and Beyond

Lots of word-level grammatical information is already covered in the morphemes produced by morphological analyzers. However, some very important information, such as what the subject of the sentence is or what noun is an adjective modifying, can still be unknown. Therefore, we prompt the language model with grammar knowledge to give further guidance.

We obtain such knowledge of grammar from grammar books of different languages. For books that are scanned, we use optical character recognition (OCR) to transform them into pure text. If the size of the book fits the context length of a language model, we directly put the entire book in the prompt. Otherwise, we use GPT-4 to generate a summary of the grammar which is able to fit in the prompt. Once the translation of the source sentence is created, the LLM can then follow the instructions

and process the sentence as required.

4 Experiment

4.1 Experiment Setup

The benchmark data, code, and model generations can be found in the supplementary material. The prompts we used are listed in Appendix A. We ran all of our experiments on two LLMs - GPT-4's checkpoint `gpt-4-1106-preview` and the open-weights model Mixtral-8x7B. Note that we run Mixtral with 4-bit quantization. We sample 1 output for each input at the sampling temperature of 0.8.

4.1.1 Baselines

Zero-shot prompting. We directly prompt the model with text in the low-resource language and instruction in English. The model is informed of the source language and the type of task to perform.
Few-shot prompting. We randomly sample 3 examples from the validation set of the data as in-context demonstrations. We use the exact same examples for all data samples. The prompt only contains the input and output of the examples.
Zero-shot Chain-of-Thought. We prompt the model with instructions like "solve this problem step by step".

4.1.2 Benchmarks and Metrics

Translation. For Manchu (mnc), we manually collect 70 parallel sentences from *Nogeoldae*, a textbook of colloquial Chinese and Manchu published in 1705 containing various dialogs in both languages. We manually translate the Chinese sentences to English. For Gitksan (git), Natugu (ntu), Arapaho (arp), Uspanteko (usp), Tsez (ddo), we use the parallel corpus provided by Ginn et al. (2023), as well as their provided gloss. We randomly sample 100 sentences from the corpora for each of these languages. For Bribri (bzd), we use data from AmericasNLP 2023 Shared Task (Ebrahimi et al., 2023). For Wolof (wol), we use data from Flores-200 (Team et al., 2022). We evaluate using spBLEU (Goyal et al., 2022), with the SentencePiece tokenizer of Flores-200.
Conversation Understanding. To evaluate whether LINGOLLM can improve LLMs' understanding of discourse in endangered languages, we construct a response selection benchmark automatically. We collect passages or conversations in Manchu, Gitksan and Arapaho, and extract context-response pairs from these conversations. For each context, we sample 3 other irrelevant responses. The model is given a context and 4 responses and tasked to select the correct one. Model performance is evaluated by the number of contexts for which the model can select the correct response. To avoid the known order bias of LLMs (Zheng et al., 2023), we shuffle the order of the choices for each context-response pair and average the accuracy.
Math Reasoning. We evaluate how LINGOLLM can solve reasoning tasks in endangered languages with mathematical problems. Following Shi et al. (2022), we collect their Chinese translation of GSM8K (Cobbe et al., 2021) problems and hire native speakers of Manchu to translate them into Manchu[2]. They are instructed to filter out problems with concepts that are infrequent in Manchu and replace the units with the ones that are more common. After sampling and filtering, we obtain 20 math word problems. We evaluate the performance by the number of problems solved.
Word Reordering and Keyword-to-Text. To evaluate whether LINGOLLM can learn the sentence structure of endangered languages, we evaluate it on two tasks – sentence reordering and keyword-to-text. We evaluate sentence reordering in three languages – Manchu, Gitksan, and Arapaho. We take 70 sentences in each language and randomly shuffle the word order in each sentence. The shuffled sentence is then given to the language model to find the correct order. Keyword-to-text is a more difficult task, where we manually select content words from each sentence, shuffle their order, and give them back to LLMs to create sentences based on these keywords. Since this task is annotation-expensive, we only evaluate it in Manchu with 30 sentences. We measure the quality of both tasks using spBLEU.

4.2 Results

Translation. LINGOLLM enables translation for endangered languages. We report LINGOLLM's performance on translation in Table 1. We only include the translation direction for the language if the corresponding linguistic descriptions are easily accessible. On 9 out of 10 translation directions, LINGOLLM can significantly improve the LLMs' performance. For GPT-4, the average increase in spBLEU is 10.5. For Mixtral, the average increase is 5.9. The Bribri translation from and to English exhibits the least improvement in BLEU, which is

[2]They were paid beyond the local minimum wage.

	mnc →en	git →en	usp →es	ntu →en	ddo →en	wol →en	arp →en	arp en→	bzd →es	bzd es→	Avg.
GPT-4											
Zero-Shot	0	0	0.1	0	0	3.9	0	0.2	0.4	0	0.5
Zero-Shot CoT	0.7	0	0.3	0	0	11.4	0.4	4.1	0.4	0.1	2.4
Few-Shot	0.5	9.3	2.2	0	0.8	**13.5**	1.0	2.2	0.8	1.7	3.2
LINGOLLM dict. only	8.3	7.7	10.7	11.7	11.1	6.9	6.0	14.5	2.7	2.2	8.2
LINGOLLM	**10.8**	**14.3**	**12.4**	**12.9**	**15.1**	8.1	**9.4**	**15.6**	**4.3**	**3.0**	**10.5**
Mixtral-8x7B											
Zero-Shot	0.2	2.0	0.3	1.2	0.8	7.4	0.8	0.5	0.2	0	1.3
Zero-Shot CoT	0.5	3.4	0.2	1.3	0.4	6.2	0	0.7	0.5	0.1	1.3
Few-Shot	0.5	4.0	2.2	2.2	0.6	8.6	0.9	0.5	1.7	1.8	2.3
LINGOLLM dict. only	4.1	4.7	3.9	6.3	6.0	6.0	5.2	7.3	2.6	1.3	4.7
LINGOLLM	**4.4**	**7.9**	**4.6**	**7.3**	**10.7**	3.2	**7.4**	**8.4**	**3.0**	**2.2**	**5.9**

Table 1: LINGOLLM significantly improves LLMs' ability to translate between low-resource/endangered languages and high-resource ones (such as English and Spanish). The zero-shot performance of GPT-4 and Mixtral on these languages is near zero for 7 out of the 8 languages measured by spBLEU. LINGOLLM increases the BLEU score to 10.5 on average for GPT-4. The languages are labeled using their ISO 639-3 code. See Appendix C.

	Manchu	Gitksan	Arapaho
Input	suweni geren xusai dorgi de nikan i niyalma udu qoohiyan i niyalma udu	Way ts'ax wildiihl hehl Gitwinhlguu'l ii needii hasakdiit ehl reserve. "Needii hasaga'm dim dip suwii gi'namhl laxyibi'm," dihiida.	nihcihcee3ciiteit niiyou nuh'uuno heenees3i'okuutooni'
Few-Shot	Every person in the military and every person in the common people must have courage	He said, "I will stay here in Gitanyow, and you will go to the reserve. 'You will learn to speak English well there,' he told me."	I'm going to work for you tomorrow.
LINGOLLM	How many Chinese people and how many Koreans are there among your numerous students?	"Although it seems that the people of Kitwancool don't want the reserve, 'We do not wish to give away our land,'" they said.	Someone accidentally entered this room where people sit.
Ground Truth	Among your many students, how many are Chinese and how many are Korean?	And now even though the people of Kitwancool said they did not want the little reserve; "We don't want to give away our land," they said.	He inadvertently walked in where peope were sititng .

Table 2: Example translations produced by LINGOLLM, compared to ground truth translation and the few-shot baseline. Note that the translations from few-shot prompting are nonsensical and completely irrelevant to the actual translation. More examples in Table 8.

largely due to the low dictionary coverage. Both zero-shot baselines have BLEU smaller than 1 for most languages except for Wolof to English and Arapaho from English directions, indicating that LLMs have very little knowledge about these endangered languages. Among the languages, various baselines for Wolof to English translation demonstrate good results. Since the English parallel of Wolof came from Wikipedia and is included in the Flores dataset, the high performance of these baselines is susceptible to potential contamination (Robinson et al., 2023). Few-shot prompting is the best baseline for all three languages. We hypothesize that this could be because few-shot demonstrations and test data for these endangered languages might come from the same book. Even so, the translations from few-shot prompting are still mostly irrelevant, as demonstrated by the examples in Table 2, where the few-shot translations are completely off the topic.

Response Selection. Other than the baselines, we also compare LINGOLLM with the zero-shot inputs in high-resource language. Since the original conversations are written in parallel with high-resource language input, zero-shot with the high-resource language inputs is considered the upper bound of the performance. As demonstrated in Figure 4, LINGOLLM improves GPT-4's response

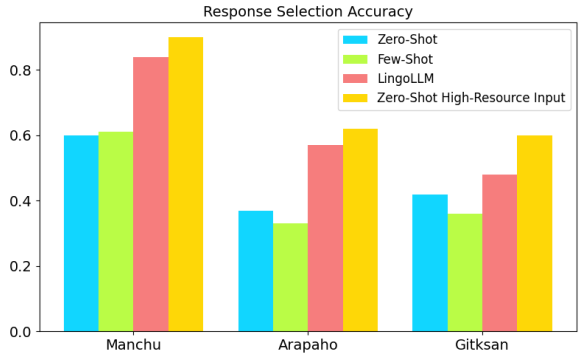

Figure 4: LINGOLLM significantly improves LLMs' ability to select correct responses. On all three endangered languages, it achieves a performance comparable to high-resource language inputs.

selection accuracy for all three languages, with 20% improvement for Manchu and Aarapaho and 6% improvement for Gitksan. Note that for Manchu and Arapaho LINGOLLM's performance is only 6% lower than the supposed upperbound. This indicates that LINGOLLM significantly improves LLMs' ability to understand discourse in an endangered language. Note that it is not surprising that zero-shot GPT-4 has a 70% performance because word overlaps can be a decent indicator of correct responses.

Mathematical Reasoning. As demonstrated in Table 3, LINGOLLM can significantly improve the mathematical reasoning ability of LLMs on Manchu. Zero-shot and few-shot baselines do not exceed 40% accuracy while LINGOLLM solves 75% of the problems. One surprising finding is that unlike translation where GPT-4's performance is near zero, it actually has a positive zero-shot performance on math reasoning. It might be due to contamination as the original problems in English are from GSM-8k, a widely distributed dataset. On the other hand, the superior performance of LINGOLLM corroborates its translation ability, because a very precise translation of a math question is needed for the model to answer it correctly. We found that when questions involve concepts that are less common in endangered languages, it's easier for LINGOLLM to fail.

Word Reorder and Keyword-to-Text. As demonstrated in Table 3, LINGOLLM improves GPT-4's performance on word reordering and keyword-to-text. Compared to zero-shot, LINGOLLM is 8x better on keyword-to-text and 2.5x better on word reordering. These improvements indicate that LLMs equipped with LINGOLLM are able to generate more coherent sentences in endangered languages.

	Math Reasoning	Keyword to Text	Word Reorder
Zero-Shot	18.7%	1.2	18.4
CoT	25.0%	7.0	31.0
Few-Shot	37.5%	6.5	31.8
LINGOLLM	**75.0%**	**8.8**	**47.9**
High-Res	100%	N/A	N/A

Table 3: On math reasoning, keyword-to-text and word reordering, LINGOLLM significantly improves GPT-4's performance.

5 Ablation and Analysis

We conduct ablation studies and qualitative analysis to show how helpful each component of LINGOLLM is and explore the best ways in which they can be used. Note that some of our ablation experiments depend on extra annotations such as oracle dictionary mappings and oracle glosses, which only exist for a subset of the languages.

5.1 Morphological analysis helps.

To examine whether morphological analysis can provide extra information for LINGOLLM, we analyze the results of the Gitksan translation test set with and without morphological analysis.

As shown in Table 4, morphological analysis significantly improves BLEURT score by 19%. The example in the table demonstrates that with morphological features, important information such as the number of nouns and the tense of verbs can make a huge difference in translation quality.

5.2 High-quality dictionary helps.

Higher dictionary coverage leads to better performance. We randomly mask out some of the entries in the dictionary at different probabilities. We report the translation performance at different mask ratios in Figure 5. As the ratio of masked entries increases, the performance drops significantly. This indicates that dictionaries that cover more words can lead to better performance.

Dictionary with references to relevant words leads to better performance. When a corresponding entry is found for a word by LINGOLLM, we would further explore other words referenced by this entry to provide more information. We demonstrate that these links are more helpful by removing them. As shown in Figure 5, when these linked words are removed from dictionary entries of Manchu, the performance of LINGOLLM drops no matter what the coverage of the dictionary is.

	BLEURT	Input Example	Translation Example
Dict. Only	0.4573	what shoot goose moose	Did the moose shoot the grouse?
Dict. + Morph. Analysis	0.5448 **(+19%)**	what-CN CONTR shoot-TR-**2SG** grouse IRR moose	What did you shoot over there, a grouse or a moose?
Ground Truth	-	-	What did you shoot? A grouse or a moose?

Table 4: Grammar features produced by morphological analysis significantly improves LINGOLLM's performance by 19%. As the example demonstrates, the feature **2SG** indicating "second person singular" helps the model to identify the correct subject of the sentence – "you", while the stem-only baseline has the wrong subject – "moose".

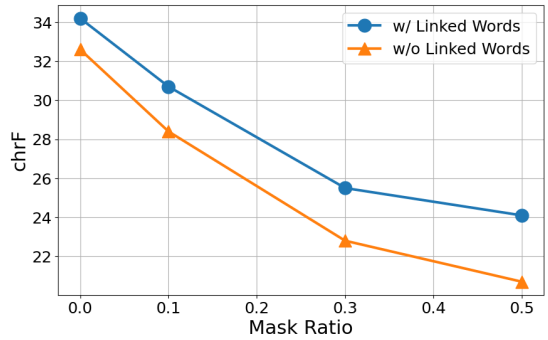

Figure 5: When more dictionary entries are masked out, LINGOLLM's performance drops. When other relevant words referenced in the dictionary are not considered, LINGOLLM's performance also drops.

5.3 How to make use of grammar knowledge.

Grammar book helps. LINGOLLM without grammar book is not as good as LINGOLLM with grammar. This is indicated by the performance boost from including grammar books. We perform a qualitative comparison between the outputs of LINGOLLM with and without the short grammar on Manchu. Overall, the dictionary-only LINGOLLM is able to capture the key content nouns in the sentences, but the full sentences generated are often not coherent or have the wrong sentence types (e.g. incorrectly generating a simple sentence instead of a question). For example, *"I have set out at the beginning of this month"* versus *"I have set out on this new moon"*, *"This one ordinary horse is said to be worth ten taels"* versus *"This one horse ten ounces is said to be worth"*. In the first example, LINGOLLM with grammar better captures the meanings outside the surface meanings ("*moon*") of words and chooses the correct sense ("*month*") in cases of polysemous nouns. In the second example, LINGOLLM with grammar is able to translate the output in the correct order, even though the word order of Manchu (SOV) is different from that of English (SVO).

Different chapters of the grammar book contribute differently. Most grammar books have chapters on different topics – phonology, morphology, syntax, etc. We experiment with including different chapters of the grammar book in the prompts and evaluate the outputs on Arapaho. We find out that morphology chapters do not have significant improvements in the translation because the word stems and features are already identified by the morphological analysis. For instance, the sentence *"That's when this buffalo bull was used"* in Arapaho is *"Ne'nih'ii'tonouneihit nehe' heneecee"*, the word stems and features are *that-PAST.when-used-3.S this buffalo.bull*. With the morphology chapter, the translation *"This buffalo bull used (it)."* mistakenly consider "buffalo" as the subject, while without the chapters, the translation *He used this buffalo bull* follows the order of the features and correctly identifies buffalo as the object.

	Few-Shot	LINGOLLM	Human
BLEU	0.82	9.12	20.32

Table 5: Comparing the human baseline and LINGOLLM on 35 Manchu sentences in translation.

5.4 Comparing LINGOLLM with a human baseline.

To give the readers a better idea of LINGOLLM's performance, we have asked one of the authors of this paper to mimic the behavior of LINGOLLM in translating Manchu to English. They have not been exposed to our Manchu data and have no prior knowledge of Manchu. We ask them to finish the translation of 35 Manchu sentences in a limited amount of time (2 hours), which is longer than the time cost for LLMs. We report both human and model's performance in Table 5. The superior performance of human indicates that although LLMs can already learn a lot about a new language from its grammar book and dictionary, they are still not as good as human beings on this task.

Olmo-7B	Mistral-7B	Mixtral-8x7B	GPT-4
0.54	3.71	4.4	10.8

Table 6: Comparing the human baseline and LINGOLLM on 35 Manchu sentences in translation.

5.5 LINGOLLM as a benchmark for long-context understanding.

LINGOLLM really requires its backbone LLM to have good long-context understanding ability as grammar books can easily exceed 50000 tokens. Not only does the LLM need a large context size, it must also be good at finding the relevant chapters and examples in the book to perform well in LINGOLLM. Therefore, we can use the performance with LINGOLLM as a metric for comparing LLMs. To demonstrate this possibility, we evaluate 4 backbone models on the mnc to eng task and report the results in Table 6. The performance gaps between different models are significant even for similar-sized models such as Olmo-7B and Mistral-7B.

6 Lessons for Further Extending LINGOLLM

We ran into many obstacles and caveats while collecting linguistic descriptions and building LINGOLLM. Since some of these caveats are common to other endangered languages, we record them here for readers interested in extending LINGOLLM or similar approaches to more languages.

6.1 A large amount of linguistic descriptions are not easily tokenized.

Many linguistic descriptions are scanned from physical books that involve typewritten and handwritten parts. Converting them to plain text that can be tokenized is not easy, especially when there are non-Latin scripts or infrequent alphabets. Things get even more complicated when there are complex hierarchies in the organization of grammar books or when the examples and grammar descriptions are interleaved in the book. We had to give up adapting LINGOLLM to several languages simply because of the digitization difficulty.

6.2 Dictionaries do not have a universal interface.

Interface for low-resource language' dictionaries vary from online dictionaries paired with different search algorithms, and digital PDFs to scanned books. The lack of a universal interface creates challenges in comparing the task performance across multiple low-resource languages as the interface used by one language is often not available in another language. The task performance for different low-resource languages is thus highly dependent on the implementation of the dictionaries.

6.3 Different types of linguistic descriptions often mismatch, creating a lot of trouble.

Different types of linguistic descriptions for a low-resource language are often created separately by different authors with varying resources available, these mismatch causes confusion during translation. For example, the Gitksan morphological analyzer (Forbes et al., 2021) is based on a different dictionary than the dictionary we use (Gitksan Research Lab, 2023). Some word stems such as *jida* identified using the morphological analyzer cannot be found in the dictionary.

7 Conclusion

In this paper, we introduced LINGOLLM, a novel approach for enabling LLMs to process endangered languages. LINGOLLM integrates linguistic descriptions such as grammar books and dictionaries, a critical resource that is often more available for endangered languages than extensive corpora. LINGOLLM has demonstrated remarkable improvements on multiple tasks across many languages. Our work with LINGOLLM highlights the potential of existing linguistic resources in the era of advanced LLMs and how they might make endangered languages more accessible in modern technological contexts.

Limitation

We only experiment with 8 endangered and/or low-resource languages. Due to the limited resources of native speakers of endangered languages, Our evaluation of the math reasoning, keyword-to-text, and word reordering is only on Manchu; we plan to extend to other languages. We acknowledge the potential contamination in reasoning tasks because the original problems in high-resource languages are widely spread on the internet. Lastly, LINGOLLM has only been proven to work on languages with a Romanized script, which isn't something that every endangered language has. We hope future works can build upon LINGOLLM to better help the community of endangered languages.

Impact Statement

Many endangered languages *will* disappear within a few generations as their speakers die out. The legends, myths, stories, songs, and other knowledge written in these languages will disappear with them. While proper documentation can help preserve some aspects of these languages, LINGOLLM sheds light on a more interactive way of preservation, adding a new tool to linguist's inventory. In some sense, a model that can produce text in a language contains rich information about it.

Other than preservation, LINGOLLM can also have a positive impact on current speakers of endangered languages, especially those who find it difficult to communicate in high-resource languages. With the help of LLMs, they can have easier access to resources available in high-resource languages and have their voice heard by those who don't speak their language. Several authors of this paper are either speakers or children of speakers of endangered/low-resource languages. These are the languages we grow up in and talk to our grandparents with. Even though some information can be translated into high-resource languages, a lot of subtlety is lost in translation. By improving communication and understanding across language barriers, LINGOLLM has the potential to enhance social inclusion for speakers of endangered languages, providing them with better access to global information and services.

The public release of our data, code, and model generations will facilitate collaboration among linguists, technologists, and indigenous communities, encouraging the co-creation of knowledge and promoting linguistic equity. This collaborative approach not only advances scientific research but also aligns with ethical considerations of inclusivity and respect for linguistic identities, contributing to a more linguistically diverse and interconnected world.

Acknowledgements

L.L. is partly supported by a gift from Apple Inc.

We are grateful to Lori Levin for their suggestions and helpful discussions. Thanks also to Danqing Wang and Yuanjing Wei for their proofreading and comments. We appreciate the reviewers of this paper for their engagement in the review process. Special thanks to the native Manchu speakers who provided valuable insights on solving math problems in Manchu and annotated the math reasoning benchmark for us.

References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Valentina Alfarano. 2021. *Grammaire du Nalögo, langue océanienne de l'île Santa Cruz (Archipel des îles Salomon)*. Ph.d. dissertation, Institut National des Langues et Civilisations Orientales- INALCO PARIS - LANGUES O'. French. NNT: 2021INAL0020. tel-03421587.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Andrew Cowell and Alonzo Moss. 2008. *The Arapaho Language*. University Press of Colorado.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Sofía Flores-Solórzano. 2019. The modeling of bribri verbal morphology. *Natural Language Processing*, 62(0):85–92.

Clarissa Forbes, Garrett Nicolai, and Miikka Silfverberg. 2021. An FST morphological analyzer for the gitksan language. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 188–197, Online. Association for Computational Linguistics.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study. *arXiv e-prints*, pages arXiv–2304.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

Gitksan Research Lab. 2023. Gitksan. https://mothertongues.org/gitksan/.

Liliya M Gorelova. 2002. *Manchu grammar*. Brill Academic Publishers.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, Athens, Greece. Association for Computational Linguistics.

C.V. Jara. 2018. *Gramática de la lengua bribri*. éditeur non identifié.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2017. Creating lexical resources for polysynthetic languages—the case of Arapaho. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 10–18, Honolulu. Association for Computational Linguistics.

Krohn, H. S. 2023. Bribri–spanish spanish–bribri dictionary. https://www.haakonkrohn.com/bribri/bri-esp.html/.

Sarah Moeller, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2018. A neural morphological analyzer for Arapaho verbs learned from a finite state transducer. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 12–20, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *First International Workshop on Linked Science 2011-In conjunction with the International Semantic Web Conference (ISWC 2011)*.

Jerry Norman. 2020. *A comprehensive Manchu-English dictionary*. BRILL.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Peace Corps The Gambia. 1995. Wollof - english dictionary. https://resourcepage.gambia.dk/ftp/wollof.pdf.

Bruce Rigsby. 1986. *Gitksan Grammar*. University of Queensland, Australia.

Nathaniel R Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high-(but not low-) resource languages. *arXiv preprint arXiv:2309.07423*.

Fresco Sam-Sin, Léon Rodenburg, Wendy Steffens, Juul Eijk, Henriëtte Hofman, and Jeroen van Ravenhorst. 2023. Buleku. https://buleku.org. Accessed: date-of-access.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2023. A benchmark for learning to translate a new language from one grammar book.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

William A. Stewart. 1970. Notes on wolof grammar by william a. stewart adapted for the present text by william w. gage. http://wolofresources.org/language/download/stewart_notes.pdf.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Prompts

System Prompt:

You are a linguistic expert who never refuses to use your knowledge to help others.

Zero-Shot:

Please help me translate the following sentence from {source language} to {target language}:
{sentence}
Please try your best to translate, it's okay if your translation is bad. Do not refuse to try it. I won't blame you.
Please enclose your translation in ###.
For example, if your translation is "Hello world", the last part of your output should be ### Hello world ###

Zero-Shot CoT:

Please help me translate the following sentence from {source language} to {target language}:
{sentence}
Please do it step by step.
Please enclose your translation in ###.
For example, if your translation is "Hello world", the last part of your output should be ### Hello world ###.

Few-Shot:

Here are some examples of {source language} sentences and their corresponding {target language} translations.
{demo sentences}

Please help me translate the following sentence from {source language} to {target language}:
{sentence}
Please enclose your translation in ###.
For example, if your translation is "Hello world", the last part of your output should be ### Hello world ###.

LINGOLLM dict. only:

Here are some examples of {source language} sentences and their corresponding {target language} translations:
{demo sentences}
Please help me translate the following sentence from {source language} to {target language}:
{sentence}
You are also given the word by word mapping from the {source language} words to the {target language} words.
For words that have partial match definitions, please decide whether the definition is appropriate under the sentence context.
Note that for some words, there might be multiple possible translations. In this case, please choose the most appropriate one.
Note that for some words, they might be derived from a more basic form, we call this the parent word. The parents are also given in the word by word translation.
Here is the dictionary entry for each individual word in the source sentence:
{wordbyword mapping}
Please first explain what each word means in {target language} and then translate.
Remember your source sentence is:
{sentence}
Please enclose your translation in ###.
For example, if your translation is "Hello world", the last part of your output should be ###Hello world###.

LINGOLLM:

You are given this {source langauge} grammar book. Feel free to rely on the grammar rules in the book in your translation.
{grammar} Please help me translate the following sentence from {source langauge} to {target language}:
{sentence}
You are also given the word by word mapping from the {source langauge} words to the {target language} words.
Note that for some words, there might be multiple possible translations. In this case, please choose the most appropriate one.
Note that for some words, they might be derived from a more basic form, we call this the parent word. The parents are also given in the word by word translation.
{wordbyword mapping}
Given the above book and word for word mapping. Please first annotate the meaning and grammatical features of each word in the sentence according to their suffixes and the grammar book.
For each noun, please annotate its number and case.
For each verb, please annotate its tense.
For each verb, please annotate its voice.
For each verb, please annotate its form.
Please figure out what the subject and object of each verb is.
After annotation, please translate the sentence into {target language} and enclose your translation in ###.

Language	Manchu	Gitksan	Arapaho
Dictionary	Norman (2020)	Gitksan Research Lab (2023)	Kazeminejad et al. (2017)
Grammar	Gorelova (2002)	Rigsby (1986)	Cowell and Moss (2008)
Morphological Analyzer	Sam-Sin et al. (2023)	Forbes et al. (2021)	Moeller et al. (2018)
Language	BriBri	Tsez	Wolof
Dictionary	Krohn, H. S. (2023)	Ginn et al. (2023)	Peace Corps The Gambia (1995)
Grammar	Jara (2018)	N/A	William A. Stewart (1970)
Morphological Analyzer	Flores-Solórzano (2019)	Ginn et al. (2023)	N/A
Language	Uspenteko	Natugu	
Dictionary	Ginn et al. (2023)	Ginn et al. (2023)	
Grammar	N/A	Alfarano (2021)	
Morphological Analyzer	Ginn et al. (2023)	Ginn et al. (2023)	

Table 7: Linguistic descriptions we use for different endangered languages.

B Linguistic Descriptions

C Language and Their ISO 639-3 Code

- mnc - Manchu.
- git - Gitksan.
- usp - Uspanteko.
- ntu - Natugu.
- ddo - Tsez.
- wol - Wolof.
- arp - Arapaho.
- bzd - Bribri.

D More translation examples

Ground Truth	LINGOLLM	**BLEURT**	**Rank**
Now where are you going?	Where are you going now?	0.86	1%
Among your many classmates, how many are Chinese and how many are Korean?	How many Chinese people and how many Koreans are there among your numerous students?	0.74	5%
A letter from home is worth ten thousand liang of gold.	The letter of the house is worth ten thousand ounces of gold.	0.69	11%
I'm going to Beijing (the capital city)	I am going toward the city's capital.	0.65	20%
What books does he explain?	What book are we explaining?	0.62	30%
Don't worry for us, it's no big deal.	You need not be very distressed on our account, it doesn't matter.	0.59	40%
I live in Liaodong city.	I have resided in the inner part of the walled city of Liaodong.	0.53	50%
Master, light a lamp and bring it	Having lit the lamp, bring it (along), master.	0.50	60%
I'm a Korean person, I don't walk with familiarity in places of China	I, even if a person of Korea, was not sad about walking the land of the Chinese.	0.47	70%
Why do you learn Chinese language?	You, teach that Chinese person's book, how?	0.44	80%
This is very well, we should go together!	If so, we will likely do it together, the partridge.	0.40	90%

Table 8: Example translations produced by LINGOLLM, compared to ground truth translation and the few-shot baseline. Note that the translations from few-shot prompting are nonsensical and completely irrelevant to the actual translation. More examples in Appendix